TU Delft OPEN

JHTR

Journal of Human-Technology Relations

ARTIFICIAL INTELLIGENCE: PANACEA OR NON-INTENTIONAL DEHUMANISATION?

Luc van der Gun luc.vandergun@ru.nl
School of Artificial Intelligence, Radboud University, Netherlands & School of Philosophy, Radboud University, Netherlands https://orcid.org/0009-0003-7635-6898

Olivia Guest olivia.guest@donders.ru.nl
Department of Cognitive Science and Artificial Intelligence, Radboud University, Netherlands & Donders Institute for Brain, Cognition & Behaviour, Radboud University, Netherlands
https://orcid.org/0000-0002-1891-0972

Article type: Student essay
Review process: Editorial review

DOI: https://doi.org/10.59490/jhtr.2024.2.7272
ISSN: 2773-2266
Submitted: 21 November 2023 Revised: 12 March 2024 Accepted: 16 April 2024 Published: 28 May 2024

How to cite (APA): Van der Gun, L., Guest, O. (2024). Artificial Intelligence: Panacea or Non-Intentional Dehumanisation?. *Journal of Human-Technology Relations*, *2*(1), pp.1-11.
https://doi.org/10.59490/jhtr.2024.2.7272

Corresponding author: Luc van der Gun

Keywords

Artificial intelligence; Technology; Enframing; Non-intentional dehumanization; Accidental consequences

Abstract

Applications of artificial intelligence (AI) are able to optimise our lives par excellence, and it is clear that this will only increase as time passes. In many ways, this is very promising, but the forms that AI takes in our society have also sparked many concerns about dehumanisation. What is often recognised is that AI systems implicitly exert social power relations—whether intentionally or not, as might be the case with bias—such that the danger would be gone if only we improved our models and uncovered this hidden realm of intentional oppression. However, these views overlook the possibility that detrimental consequences may also arise precisely because AI is able to attain favourable goals flawlessly. This problem of adverse side effects, which are strictly accidental to the goals we set for AI to effectuate, is explored through the notion of "non-intentional dehumanisation". To articulate this phenomenon, this essay consists of two parts. The first part will establish how naive AI usage presents a paradigmatic case of this problem. In the second part, we will argue that these issues occur in a two-fold fashion; not only does AI risk inducing harm to the "used-upon", but also to the user. It is with this conceptual model that awareness may be brought to the counter side of our ready acceptance of AI solutions.

1 INTRODUCTION

In recent years, artificial intelligence (AI) has become a central topic of public debate. On the one hand, it is framed as a "silver bullet" to alleviate us from our labour and complex social problems (Leufer, 2020). On the other hand, great concerns are also expressed, in that AI can be a means to oppress people (Erscoi et al., 2023; Kerr, 2021) or because it stands at risk of moving beyond our control (such as with the paperclip maximiser, see e.g. Murphy, 2018). The present essay joins with this side of worry, for it is a response to the "panacea" (a "cure-all") view of AI that some of its proponents advertise. In this, it maintains that a crucial aspect of this danger has remained under-discussed. While the potential of AI for *intentional* dehumanisation has received recognition as of late, its *non-intentional* counterpart has not yet attracted much methodical treatment.

To illustrate this difference, two contemporary artificial facial feature-detection systems may serve as examples. In the case of AI-enabled persecution of ethnic minorities (e.g., Byler, 2019), we often judge that the dehumanisation of people with specific racial characteristics is indeed intentional. When AI is used to predict high-risk travellers for border-control purposes (e.g., Biddle, 2018), however, the underlying intentions may be classified as "good", but negative consequences might still be at hand simply by virtue of a high rate of false positives. In this latter case, the harm done is a side-effect rather than the goal.

As such, *non-intentional dehumanisation* is strictly *accidental* to attaining certain goals, which may in themselves be desirable. It expresses that implicit consequences might be at hand even when these goals are perfectly reached. This removes the focus from harms that stem from algorithmic inaccuracies, such as certain forms of bias. In these cases, the goal was not reached in the first place, and an improvement to the quality of the data may already be sufficient to alleviate the problem at hand. The concept of non-intentional dehumanisation aims to recognise the fact that even data without “mistakes” put at work to achieve favourable goals may induce harm.

To get a clear view on this phenomenon, this essay is divided into two parts. Firstly, the possibility of non-intentional consequences is explored through Heidegger’s non-neutral “essence” of technology. This allows us to characterise the relation between its occurrence and the reductive aspect of contemporary AI systems. In the second part, it will be elaborated that non-intentional dehumanisation has a two-fold form; not only should we attend to its effects on the side of the “used-upon”, but, unconventionally, also on the side of the user. These forms are identified as a “denial" and “deprivation" of humanness, respectively. With this conception of non-intentional accidental consequences, light can be shed on the harmful effects that remain hidden under the veil of a positive appreciation of the ability of AI to accomplish our goals.

2 THE RELATION BETWEEN AI AND REDUCTION

2.1 DISENTANGLING NEUTRALITY AND CONTROL

To conceptualise how and why non-intentional dehumanisation emerges through our AI usage, we must ask the following question: How is it possible that a perfectly functioning technological tool could induce harmful side effects, even when used to attain a favourable goal? To conceive how side effects and usage can co-occur, we must first find a way to differentiate the neutrality of technology from our control over it. In fact, the assimilation between the two is the principal reason as to why non-intentional harms tend to fly under the radar.

One such attempt to dissociate the two can be seen in Pfaffenberger’s (1988) account against two common antagonistic understandings of technology, in which it is held to be either neutral and controllable or non-neutral and uncontrollable. He argues that both views overlook the hidden social relations underlying technology usage. Technology should rather be seen as non-neutral precisely because it enacts our oppressive human control. A similar intentional view of technology can also be seen in Munn’s (2022) account against the “myth of automation”, where he states that this great promise conceals its socio-political interests. Yet, if one takes their argument to its logical conclusion, this non-neutrality quickly disappears because the discussed dehumanising consequences are linked to algorithmic mistakes or abusive human control. Implicitly, technology is still conceived as a neutral extension of our actions, and the truly accidental-non-intentional side of our technology usage remains glossed over.

To understand how we may fully disentangle neutrality from control, we may turn to Heidegger. For him, technology is non-neutral because it is a sort of lens that (unnoticeably) shapes how the world is presented to us. This way, the “essence” of technology presents a paradigm of efficiency and reduction. From such a view, we can see that negative consequences may emerge even if we are in control and have solely “good” intentions behind deploying our digital instruments.

2.2 HEIDEGGER’S ESSENCE OF TECHNOLOGY: ENFRAMING

In *The Question Concerning Technology*, Heidegger (1977, published in 1954 in German) attacked our everyday “instrumental definition” of technology. Technology is not simply a neutral means to an end employed in human activities, but must instead be characterised by a more mysterious “revealing”. This revealing can be seen in the way that a hammer may be used to build a wooden bridge, or when a smartphone is used to access novel information so long as one has an internet connection. For Heidegger, technology allows us to see things that were not naturally there; it is a way of “revealing” something. Although we remain in control, this revealing is beyond mere human doing since we do not give rise to what is revealed solely by ourselves. The non-neutrality of technology then lies in the fact that it is not at all evident what effects it brings about.

Now, Heidegger claims that it is especially with modern technology that this revealing takes a turn for the worse; the original, more innocent, and creative type of revealing suddenly unfolds itself into a more dominating form. This has happened, he argues, because of its connection to modern science, which strives to render the world fully calculable and predictable. Consequently, the world and other human beings are reframed as being nothing but static resources ("standing-reserves") to be challenged into existence and commanded around. He illustrates this turn-around with a striking example: Whereas before, we let nature take care of our farming—tools were used only to set the growth of crops in motion—we now challenge a plot of land to produce the same crops over and over, thereby damaging the soil.

This leads Heidegger to conceive of a problematic essence of technology, which he calls "enframing". Modern technology testifies to the existence of a "framework" of our understanding through which the world is made visible as being nothing but exploitable resources. In this way, his essence of technology is not simply a description that exhaustively characterises all technological artefacts. Enframing pertains to how we bring technology in relation to the world, such that it has more to do with the fact that technology is so omnipresent in our contemporary society than that these technologies might be problematic in themselves. In other words, the problem is that we lose ourselves to the instrumental definition of conceiving technology neutrally, such that the technological mindset entrances us, as it were.

What we may draw from this conception of enframing is that it reshapes the way in which we think about the consequences of technology usage; perhaps it is not evil intent that poses the biggest threat, but rather that we are set up to perceive and act in an entirely reducing way, all while thinking that we know what we are doing. Our simple strive for efficiency, something which is indeed not problematic in itself, can be seen to give rise to disastrous effects nonetheless. However, while for Heidegger, technology is an all-pervasive (ontological) account of how modern society perceives reality, we want to add that with AI, this enframing has, in some way, become materialised in concrete artefacts. Similar to how we employ technology in a challenging way, AI is made to do the same *for* us.

2.3 AI: DATAFICATION AND OPTIMISATION

Now that we have seen how the mindset of efficiency is inherently related to reduction, we can take the step toward AI applications in particular. Needless to say, it is invariably difficult—if not impossible—to characterise its variety of forms under one denominator. However, AI systems do share certain commonalities, or tendencies, when it comes to their constitution and their capabilities.

AI, broadly conceived, is different from other kinds of technology in that it strives to learn something *for* us. As Heidegger has argued that we must take distance from the machine if we are to avoid machine-like thinking, artificial intelligence—the very idea of which seems to be mechanised thinking—would be a recipe for disaster. As such, our claim is that AI is not only employed in this paradigm of efficiency but that it is also a crystallisation of it.

To see this, we must turn to the interplay of datafication and optimisation, a process of quantification and subsequent comparison, which is visible in the inner workings of AI algorithms, our development of them, and our deployment of these systems in the world. Regarding the first, AI algorithms—e.g., reinforcement learning or deep learning—are at the root governed by an optimisation function to achieve the desired accuracy. One only needs to look at large language models, personalised news feeds and artwork generators to see that their output is some maximum with respect to their input. This optimisation builds on datafication; before items can be compared, the world must first be reduced to its "interesting parts". In this sense, reduction is apparent in the very constitution of AI, precisely in maximising (challenging) the presence of one aspect while, and by means of, disregarding all others.

This dynamic of datafication and optimisation is also visible in how we develop AI systems. As Birhane et al. (2022) have shown in their systematic literature review, the "state-of-the-art" label—which relates to non-human technical values—dominates the contemporary research scene, which suggests that it has become the sole legitimisation for research. Similarly, Zawieska (2020) speaks of an "engineering ethos" in the field of robotics. She argues that technical functionality supersedes a more human-oriented approach such that a "tacit" dehumanisation is inherent to the current design process. Yet, however much in line this reappraisal of the living human being over the efficient and technical is with the current essay, these accounts do not yet adequately grasp that the technical mindset is also human, in that we employ it to attain certain desirable ends. Again, the problem is rather that the naive belief of the neutrality of these values of performance and efficiency—that they are dependent *only* on the goal that sets them to work—has the inherent tendency to overrule other human values in an almost completely concealed fashion.

Lastly, this same dynamic also pertains to our use of AI within the world. In this light, Vrontis et al. (2022) note that with the introduction of AI, technology can overrule previously uniquely human tasks, such as communication and interaction. Thus, in its capacity for "intelligence", AI allows for an unprecedented output comparison between humans and machines. Although the replacement of humans by AI systems appears reasonable when we look only at its increase in efficiency, this comparison conceals the elimination of other human qualities. To some extent, AI has also made the prediction of human behaviour possible, thus heightening the risk of controlling and challenging human beings (Förster, 2019). In ways such as these, the introduction of learning algorithms allows for an increased presence of this dynamic of quantification and comparison in the real world, such that we might say that Heidegger's mostly abstract diagnosis has established a material presence.

In all these domains, the specific capacities of AI can be seen to aggravate the problems that emerge from its non-neutrality. However, we may also identify characteristics of AI that fall beyond this interplay of datafication and optimisation. In this light, we may point to the increased interconnectedness that has arisen with information technologies and "smart" devices, where data from one application is passed onto the next (Stolterman & Fors, 2004). As Förster (2019) argues, AI delivers unparalleled invisibility, or opaqueness, not only in that it is a black box, but also because it is software that can be hidden and run using increasingly speedy microprocessors. This way, both interconnectedness and invisibility multiply the presence of the dynamics already at hand.

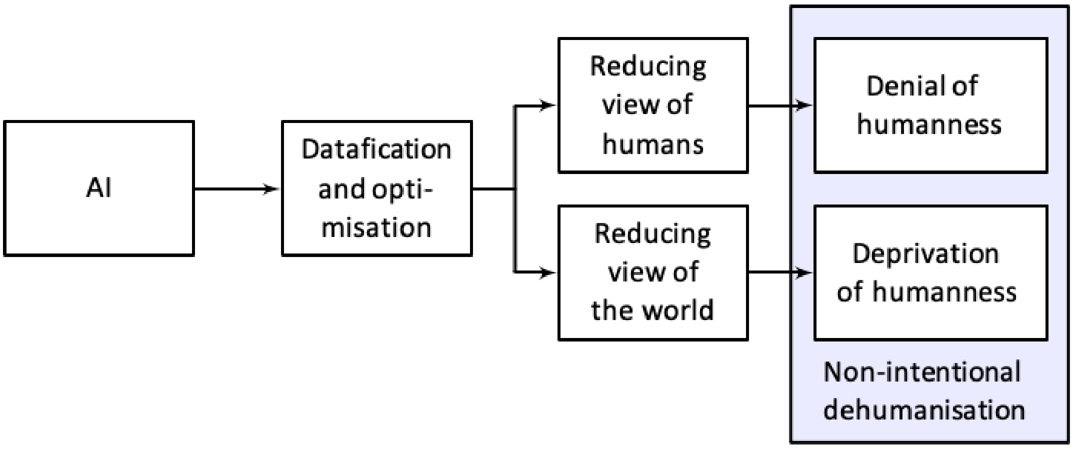
In any case, it has become clear that AI, as an embodiment of efficiency, is inherently related to reduction. While it seems as though it is a neutral means to exert and increase our control, it effectuates technology-enabled goals that are themselves already reducing. To see how this makes AI an architecture for non-intentional dehumanisation, we must describe the effects of this reduction in more detail—we turn to this in the next part.

3 A TWO-FOLD MODEL OF NON-INTENTIONAL DEHUMANISATION

3.1 A PERSPECTIVAL ACCOUNT

In the analysis of the first part, it was made apparent how AI has inherent ties to reduction. Here, we will explicate the two ways in which this reduction takes form in order to get a clear view of the specific effects that characterise non-intentional dehumanisation. Essentially, this dichotomy amounts to a difference of perspective; technology is a lens through which one can be viewed by others and simultaneously look through oneself. Not only may technology play a part in a re-description of human beings *as* being mere resources, it may also diminish our very experience of a rich and lively world. This differentiates what might be called the “used-upon” and the “user”, which are referred to as a “denial of humanness” and a “deprivation of humanness”, respectively (see Figure 1). It is a difference between *self* and *other*—who is it that utilises the AI system, and who is it that undergoes its side effects? To shed light on these phenomena, we will investigate how both forms are non-intentionally apparent with technology, particularly AI, and which specific effects they may give rise to.

Fig. 1. *A schematic which shows the part AI applications play in the onset of non-intentional dehumanisation. The effects of this phenomenon are two-fold: not only may AI lead to reduced views of other human beings (a denial of humanness), it may also directly diminish our self-conceptions and reduce our views of the world (a deprivation of humanness).*


AI
Datafication and opti-misation
Reducing view of humans
Reducing view of the world
Denial of humanness
Deprivation of humanness
Non-intentional dehumanisation

3.2 NON-INTENTIONAL DENIAL OF HUMANNESS

In his paper, Haslam (2006) identifies two separate forms of dehumanisation, both of which are the *denial* of a certain humanness. On the one hand, there is the intentional “animalistic” kind, which displays the denial of the equality or reasonableness of specific human beings to legitimise one group or person to be cruel to another. On the other hand, there is a non-intentional “mechanistic” form that denies people their aliveness, emotionality, and agency by likening them to machines. While the former version is generally recognised, he argues, the latter tends to remain unnoticed. Interestingly, and accordingly, he elaborates this latter abstraction of human beings to occur through a “cognitive bias” induced by technology—that is, the “robotic pursuit of efficiency and regularity” (Haslam, 2006, p. 253). In this way, we may characterise his dual account as the difference between a conscious (malicious) displacement of another human being and a more unconscious indifference to this “another”. Although both forms of denial are necessarily interpersonal, technology may induce non-intentional denial by virtue of its influence on how two people perceive one another.

When it comes to AI systems, this non-intentional denial of humanness can principally be seen in the act of comparing human performance to that of the algorithm, as well as in the establishment of algorithmic governance. Regarding the former, we can think of the ease of replacing humans with their algorithmic counterparts by comparing their effectivity, that is, their contribution to company revenue. With respect to the latter, the introduction of AI in the sphere of human resource management leads to more and more company decisions being made in light of maximising performance output (Vrontis et al., 2022). Consequently, human labourers doing their job under AI “supervision” are subject to increased control and a heightened distance from the management team. Because AI could advance certain company processes, humans may be disregarded—completely or psychologically—without the intent to harm them.

A similar phenomenon is apparent with the use of AI in legal adjudication processes. Here, algorithms are increasingly used to reduce costs and to increase speed and standardisation—employing their ability to circumvent human weaknesses such as bias and inconsistency (Re & Solow-Niederman, 2019). Yet, Re and Solow-Niederman elaborate that the introduction of these artificial adjudicators shows that the prosecuted are subject to additional harms that are distinct from the contents of the prosecution itself. In contrast to a merciful judge, AI algorithms make incomprehensible decisions against which no arguments can be given. As such, they argue that, paradoxically, the standardisation through AI may result in the law being viewed as less reasonable, leading people to feel disempowered and vulnerable.

Perhaps the best example can be seen in Karches' (2018) analysis of the electronic health record. The value of such AI-based medical instruments lies in the fact that it may improve healthcare by providing valuable aids to the diagnosis process. But, Karches contends, this comes at the price of a drastic depersonalisation, since medical advice is now derived from cohorts of similar demographic data, rather than from the unique individual. Continuing, he argues that the electronic health record stands in stark contrast to simpler medical tools such as the stethoscope, in that the latter allows for a closer examination of the patient's body, whereas the former leads to no such proximity at all. Thus, we see the same non-intentional phenomenon as before. Precisely in the potential of AI to enhance healthcare on a systemic level by increasing the amount of people "cured", its use implicitly results in a relatively indifferent treatment of the individual.

3.3 NON-INTENTIONAL DEPRIVATION OF HUMANNESS

The preceding account has shown how AI systems may induce accidental effects for people who relate to its implementation in terms of being "used-upon". In a certain way, the given examples have shown that the use of AI systems makes *managing* more important than what is *managed*. However, because technology can also be used alone—beyond an interpersonal context—, it is important to show how AI might also lead to accidental effects in direct relation to the "user".

To this end, Borgmann's (1987) "device paradigm" may be used to characterise a deprivation of humanness. In his book *Technology and the Character of Contemporary Life*, he stresses the importance of "focal practices" and "focal things", which impart those experiences that we consider to be part of a good (human) life. His device paradigm testifies to the dynamic where complex "things" are readily replaced by simple "devices" because the latter is made specifically to outclass the former regarding a specific and singular *functionality*. To illustrate this, he gives a powerful example: the hearth, which requires planning and attention in its use, is replaced by the central heating system, simply because it is *better* at heating. In other words, Borgmann shows how commodities such as heat are more "available" through devices, in that the heating system requires only the turn of a knob and can heat all rooms with relative ease. Yet, he argues that this paradigm comes with a counter-side: the "disburdening" character of the device leaves us "debilitated". While "things", such as the hearth, brought the practice of skills and a distribution of social roles within one's family, "devices" leave its users dependent, deskilled, and dissociated from social and bodily engagement.

Deprivation of humanness thus does not relate to people being *subjected* to AI's capacities, such as with algorithmic governance or the replacement by outperforming machines, but rather pertains to the individual's personal decisions. This description is perhaps most clearly applicable to the technology of smart homes. Borgmann stated that "the concealment of the machinery and the disburdening character of the device go hand in hand" in that "a commodity is truly available when it can be enjoyed as a mere end, unencumbered by means" (Borgmann, 1987, p. 60). Analogously, smart home applications are embedded within our homes, to provide us with comfort on command (Wilson et al., 2015). But, as Wilson et al. elaborate, this delivery of commodities extends even beyond command, since the promise of smart homes implies that these technologies are believed to be able to fulfil the user's needs most optimally, rather than the users themselves. Consequently, Wilson et al. argue, smart homes infringe upon the user's relationships, as well as their domestic roles.

Within professional spheres, deprivation also applies to the managers who decide to employ AI systems. As Selenko et al. (2022) elaborate, a management team that has replaced all human workers with more efficient machines unwittingly also deprive themselves of human interaction, as there might not be anyone left to interact with but machines. Moreover, they argue that even job augmentation by outsourcing parts of labour to the machine may reduce one's work to entail mere button-pressing, weakening one's self-esteem and sense of meaningfulness. In a similar vein, Fritts and Cabrera (2021) show that the addition of recruitment algorithms comes at the price of diminishing the employee-employer relationship, even though they might lead to quicker, more accurate, and objective judgements. They argue that in this case, it is not so much that the applicants feel denied in their humanness, but rather that they feel deprived, in that they are no longer able to truly convince anyone of their skills or personality, leading to so-called "hollow victories". This way, with its human-like figure, AI is able to replace actual human elements, thereby substituting interpersonal contexts with one of solitude.

3.4 A SYNTHESIS OF BOTH ACCOUNTS

Denial and deprivation of humanness amount to a disregard for another's experience (one's being-experienced) and the hollowing out of one's own experience, respectively. What these forms of non-intentional dehumanisation share is that they both stem from the unparalleled capacity of AI systems to reach our—in themselves, potentially humanising—goals. Yet, a differentiation of the two forms is particularly valuable when it comes to the question of responsibility; while the former case invites us to see that someone else is responsible, such that we might speak of non-intentional oppression, the latter case shows that it is the user who remains (largely) responsible. Thus, by identifying denial and deprivation of humanness *as such*, we can better take action to resolve the harms at hand. However, before we can conclude this part, a few considerations must be had.

Firstly, the use of the notion of intentionality to discriminate the phenomenon from the cases where AI does not perfectly reach its goal, or where this goal is inherently harmful as such, subjects this account to the problem that intentionality and non-intentionality may sometimes be indifferentiable. This issue pertains to the interpersonal form of denial in particular; how far does the non-intentionality stretch if someone has knowledge of their disregard for other individuals, as might be the case when human workers are replaced with machines for monetary reasons? While cases like these present a difficulty for our assessment, it remains necessary to keep the accidentality of the consequences close by. The crux lies in the fact that the datafication and optimisation process surrounding AI, precisely in its capacity to improve certain measures, makes it easier to forget about the personhood of other human beings. In this sense, technology plays an irreducible role in bringing what we have here called "non-intentional denial"—the same holds for the form of deprivation.

A second consideration to be expressed is that the harmfulness of these AI-induced side effects depends on the person and their context in question. Especially in medical care, dehumanising *means* (inflicting pain) must be weighed against the humanising *end* of curing the person in question (Dawson, 2021; Palmer & Schwan, 2022). In these cases, (accidental) harm may be a necessary evil to achieve greater goods because viewing a patient as a machine to be fixed is beneficial for both the surgeon and the person to be healed. Similarly, Wilson et al. (2015) explain that the risks of smart homes are not as worrisome when they are applied in light of providing vulnerable people (e.g., the elderly or those who are disabled) with greater independence and safety. These examples may be extended to a general point, stressing that accidental effects (non-intentional dehumanisation) are only relevant so long as the context is taken into account in its entirety. Needless to say, the line between desirable and undesirable AI applications is thin, but the identification of accidental effects can be used to bring this context of evaluation into clearer view.

In sum, the specific characteristics of AI—the dynamic of datafication and optimisation—may lead to accidental-non-intentional consequences in its context of use. The specific tendencies within AI, as well as our employment of it, are then the causes of these consequences. In the case of denial of humanness, we might note that the comparison of humans with machines, systematisation, instrumentalisation, and the black-box character of AI tend to give rise to effects in terms of disregard for feelings, depersonalisation, passification and domination. In the case of deprivation of humanness, we should rather think of availability, hiding of the means (invisible technology), and removal of a human touch to lead to effects such as unfulfillment, social isolation, deskilling, dependence, and dissociation from nature. Notably, these lists are by no means exhaustive, and are simply meant to give words to the type of properties that one could think of. Their principal use is *heuristic*, to find the as-of-yet unnoticed non-intentional consequences.

4 CONCLUSION

In this essay, we have put the notion of non-intentional dehumanisation on view, which characterises harmful consequences that are strictly accidental to the perfect attainment of desirable goals. These detriments were elaborated to stem from our ready deployment of AI and from its constitution, both of which display a dynamic of datafication and optimisation. Following from this point, the dual nature of this AI-enabled reduction of a certain humanness was explored.

On the one hand, AI relates to people in the sense of being used-upon, such that some accidental effects must be understood in terms of a denial of humanness. On the other hand, AI also has consequences for its users, for whom distinct effects which concern a deprivation of humanness may be identified. This way, naive AI usage—although this also holds for other technologies—leads us to be experienced by others in a reduced fashion, as well as reducing our own experience of our lives and the world.

This does not mean that we can disregard the (intentional) socio-political dimension of technology usage, from which the framing of non-intentional dehumanisation had separated at its conception. We do not intend for this non-intentional account of AI to overrule and dismiss intentional forms of dehumanisation that are at least equally harmful, such as bias and explicit oppression of minorities. Technology should never be seen to be the sole driving force of harmful effects—even in the case of non-intentional dehumanisation, it is we who remain responsible.

It also does not mean that we must be unapologetically critical of AI as a whole. We do not have to give up the immense advancements that it has booked in sectors such as healthcare, such as AI-enabled radiology. As Borgmann (1987) already recognised, some technologies may even enhance current focal practices or establish new ones by delivering the required time, equipment, and instruments. It is therefore necessary to remain in active discussion with relevant experts and society, not only because AI applications are so difficult to assess, but also because our current world sees AI innovations develop at an exponential pace. The conception of non-intentional consequences may then prove useful, precisely to aid in this evaluation process, in the first place to bring awareness to the accidental consequences that tend to remain invisible, but also to catch these harmful side effects early on—perhaps even while AI systems are still in development. In this sense, the general tendencies of AI outlined in this essay are to be used as a guideline to identify the possible onset of this hidden and devastating side of novel AI applications.

What stands most central to the phenomenon of non-intentional dehumanisation is that we should rigorously reflect on the goals of efficiency and performance that we hold dear, whether this be professionally or in our own homes, and what reducing consequences are inherent to these goals we set for AI to achieve. Hopefully, bringing to light this obfuscated form of dehumanisation may help us steer away from a ready acceptance of potentially harmful AI solutions, and provide us with the opportunity to select those applications that are actually valuable and virtuous.

Data Access Statement

No new data were generated or analysed during this study.

Contributor Statement

Luc van der Gun is responsible for conceptualising and writing the original draft, reviewing, and editing. Olivia Guest is responsible for the supervision and editing of this paper.

Use of AI

N/A.

Acknowledgements

The authors would like to thank the Computational Cognitive Science group for their feedback, especially Iris van Rooij, Annelies Kleinherenbrink, and Laura van de Braak.

References

Biddle, S. (2018). Homeland security will let computers predict who might be a terrorist on your plane - just don’t ask how it works. Retrieved July 4, 2023, from https://theintercept.com/2018/12/03/air-travel-surveillance-homeland-security/

Birhane, A., Kalluri, P., Card, D., Agnew, W., Dotan, R., & Bao, M. (2022). The Values Encoded in Machine Learning Research. *ACM International Conference Proceeding Series*, 173–184.

Borgmann, A. (1987). *Technology and the Character of Contemporary Life*. University of Chicago Press.

Byler, D. (2019). China’s hi-tech war on its muslim minority. Retrieved July 4, 2023, from https://www.theguardian.com/news/2019/apr/11/china-hi-tech-war-on-muslim-minority-xinjiang-uighurs-surveillance-face-recognition

Dawson, J. (2021). Life & times medically optimised: healthcare language and dehumanisation. *British Journal of General Practice, 71*(706), 224.

Erscoi, L. A., Kleinherenbrink, A., & Guest, O. (2023). *Pygmalion Displacement: When Humanising AI Dehumanises Women* [SocArXiv].

Förster, Y. (2019). Artificial Intelligence: Invisible Agencies in the Folds of Technological Cultures. In A. Sudmann (Ed.), *The democratization of artificial intelligence* (The Democr, pp. 175–188). transcript.

Fritts, M., & Cabrera, F. (2021). AI recruitment algorithms and the dehumanization problem. *Ethics and Information Technology, 23*(4), 791–801.

Haslam, N. (2006). Dehumanization: An integrative review. *Personality and Social Psychology Review, 10*(3), 252–264.

Heidegger, M. (1977). The Question Concerning Technology. In G. Vattimo & R. T. Valgenti (Eds.), *The question concerning technology and other essays* (pp. 3–35). Harper Torchbooks.

Karches, K. E. (2018). Against the iDoctor: why artificial intelligence should not replace physician judgment. *Theoretical Medicine and Bioethics, 39*(2), 91–110.

Kerr, A. D. (2021). Artificial Intelligence, Gender, and Oppression. In W. L. Filho, A. M. Azul, L. Brandli, A. L. Salvia, & T. Wall (Eds.), *Gender equality* (pp. 54–64). Springer.

Leufer, D. (2020). Why We Need to Bust Some Myths about AI. *Patterns, 1*(7), 100124.

Munn, L. (2022). *Automation is a Myth*. Stanford University Press.

Murphy, J. (2018). Artificial Intelligence, Rationality, and the World Wide Web. *IEEE Intelligent Systems, 33*(1), 98–103.

Palmer, A., & Schwan, D. (2022). Beneficent dehumanization: Employing artificial intelligence and carebots to mitigate shame-induced barriers to medical care. *Bioethics, 36*(2), 187–193.

Pfaffenberger, B. (1988). Fetishised objects and humanised nature: towards an anthropology of technology. *Man, 23*(2), 236–252.

Re, R. M., & Solow-Niederman, A. (2019). Developing Artificially Intelligent Justice. 242, 242–289.

Selenko, E., Bankins, S., Shoss, M., Warburton, J., & Restubog, S. L. D. (2022). Artificial Intelligence and the Future of Work: A Functional-Identity Perspective. *Current Directions in Psychological Science, 31*(3), 272–279.

Stolterman, E., & Fors, A. K. (2004). Information Technology and the Good Life, Information Systems Research: Relevant Theory and Informed Practice. *Information Systems Research. IFIP International Federation for Information Processing*, 687–692.

Vrontis, D., Christofi, M., Pereira, V., Tarba, S., Makrides, A., & Trichina, E. (2022). Artificial intelligence, robotics, advanced technologies and human resource management: a systematic review. *International Journal of Human Resource Management, 33*(6), 1237–1266.

Wilson, C., Hargreaves, T., & Hauxwell-Baldwin, R. (2015). Smart homes and their users: a systematic analysis and key challenges. *Personal and Ubiquitous Computing, 19*(2), 463–476.

Zawieska, K. (2020). Disengagement with ethics in robotics as a tacit form of dehumanisation. *AI and Society, 35*(4), 869–883.